用文献轮廓挖掘大肠癌转移芯片表达谱

黄仲曦袁孙 青袁丁彦青袁姚开泰渊第一军医大学病理教研室肿瘤研究所袁广东 广州 510515冤

摘要院目的 寻找新的大肠癌转移相关基因遥方法 根据大肠癌转移芯片的表达谱袁采用基于文献轮廓的数据挖掘方法袁从 Medline 文献数据库中提取基因的相关文献并分析词的频率袁再基于重复发生和共发生的过滤标准提取功能相关的词袁最后基于词的发生频率对基因进行功能聚类袁进一步结合文献及已有的分子生物学检测结果进行分析遥结果 发现两个新的可能与大肠癌转移相关的基因 TIAM1 和 NM23H1遥

关键词院大肠癌曰肿瘤转移曰基因芯片曰信息存储与检索
中图分类号院R735.3; G350 文献标识码院A 文章编号院000-2588渊003冤1-1195-03

Mining microarray gene expression data of metastatic colorectal cancer by literature profiling

HUANG Zhong-xi, SUN Qing, DING Yan-qing, YAO Kai-tai
Institute of Cancer Research, Department of Pathology, FirstMilitaryMedical University, Guangzhou, 510515, China

Abstract: Objective To search new metastatic colorectal cancer-related genes. **Method** Metastatic colorectal cancer microarray gene expression data was mined by literature profiling, which was based on the analysis of literature profiles generated by extracting the frequencies of certain terms from the abstracts in the Medline literature database. The terms are then filtered on the basis of both repetitive occurrence and co-occurrence among multiple gene entries. Clustering analysis was subsequently performed on the retained frequency values, shaping a coherent picture of the functional relationship among the heterogeneous genes identified in the long lists. The clustering result was analyzed against the result of document analysis and experiment. **Result and Conclusion** Two new genes (TIAM1 and NM23H1) with potential relation to metastatic colorectal cancer were identified.
Key words: colorectal cancer; neoplasm metastasis; microarray; information storage and retrieval

目前袁虽然人们已发现不少大肠癌的转移相关基因袁但还不能完全阐明其转移机制袁这与大肠癌转移本身的异质性和复杂性有关遥既往研究仅限于一种或几种基因的表达袁不能全面反映各基因间的关系遥

本教研室自行研制了肿瘤转移相关基因 cDNA 芯片[1]袁该芯片收集了目前已知的 399 个转移相关基因并对几个主要环节的多项技术指标进行优化袁使其灵敏度达到检测 5 mg 总 RNA袁且图像扫描信号清晰尧富有层次尧背景均匀尧重复性好遥我们应用此芯片对大肠癌细胞系及组织标本进行检测袁结果发现了与大肠癌转移密切相关的表达基因 51 个袁包括上调基因 22 个尧下调基因 29 个遥部分基因经 Northern 点杂交和 RT-PCR 半定量检测袁结果与芯片一致袁验证了该芯片的实用性和可靠性[2]遥

如何分析这 51 个基因的功能关系以及发现新的大肠癌转移相关基因呢钥为此袁我们采用一种基于文献轮廓挖掘微阵列表达数据的生物信息学方法来研究这些基因的功能关系[3]袁并发现了新的大肠癌转移相关基因遥

1 材料和方法

收稿日期院003-06-27
作者简介院黄仲曦(1974-)袁男袁福建漳浦人袁第一军医大学在读博士研究生袁电话院20-61640114-89099袁e-mail: zxhuang@fimmu.com

1.1 获取差异表达基因

从基因表达谱数据中袁以比值大于 2 和小于 0.5 为阈值渊这是根据 95%的可信区间[-1.96,1.96]定义的袁详细请参考文献[4]冤袁选取已知基因名称的高表达基因和低表达基因遥

1.2 获取基因的相关摘要

我们通过查询 PubMed 中含有基因名字或缩写或别称的条目来获取各个基因的相关文献遥关于基因命名的信息从人类基因命名委员会渊HGNC冤的网站上和 NCBI 的 Locuslink 的网站上获取 渊目前 HGNC 和 Locuslink 收录的人类基因已经超过 17 000 个冤遥由于相当大量的文献不采用官方基因名称袁而且所使用的基因名称和缩写可能会与 HGNC 及 Locuslink 提供的别名不同或缺少特异性渊例如公共词汇袁可能具有基因名称以外的其他意思冤袁尤其是短的只取首字母的缩写词袁因此一方面会存在漏检现象袁另一方面则出现假阳性 渊即文献中出现基因名称的检索词袁但它不代表基因而代表其他意思袁因此该文献与所查询的基因无关袁属假阳性冤遥为了避免过高的假阳性袁有必要快速浏览搜索结果以便发现并删除不恰当的检索字符串遥但是袁要提高查全率则必需尽量收全所有的别名和缩写袁这可以通过增加其他名称数据库渊例如基因组数据库 GDB冤来实现曰考虑到我们希望获得的是代表基因特征的高频率词袁因此只需得到大

部分与该基因相关的文献就可以获得这些词袁无需查全所有相关文献冤遥在 PUBMED 中袁首先在标题中检索那些包含基因的官方名称尧缩写或别称的条目遥如果检索到的文献不足 5 篇袁则扩展到摘要中进行检索遥如果还是不足 5 篇袁则用基因家族名称代替基因名称进行检索遥如果还是不足 5 篇袁则该基因要么不分析袁若分析则需十分谨慎遥

1.3 文字分析

选择恰当的输出格式渊默认输出格式是野摘要冶袁必须改为野XML冶冤后点击工具条的保存按钮就可以将 PubMed 的查询结果保存下来遥摘要从输出文件中抽提出来袁并保存在一个新文件中袁一篇摘要保存一行遥用 Montreal 公司的 Provalis Research 软件的 Wordstat 模块的野文本转换魔术师冶对每个文件进行格式转化遥再用 simstat 模块打开并用野内容分析冶的统计方法进行分析遥输出选择词的类别百分比遥

1.4 数据过滤

在分析的文献中发现的每一个特定词都赋予一个发生频率值袁这样每一个基因都有好几万条记录遥这些词当中大多数要么是普遍存在的 渊例如袁野a冶袁野because冶袁野cell冶袁野identified冶在大多数摘要中都出现冤要么是极罕见的袁因此对于定义基因特异性词的发生轮廓没有多大用处袁必须删除遥剩下的词则是出现在少数基因的大多数摘要中袁从而传递了这些基因的相关信息遥

过滤规则院第一步袁删除在科学文献当中常见的词遥用每个词在 250 个随机选取的基因中的发生值的平均值来确定该词的基值遥把基值超过 5%的词归为无辨别力的一类并删除遥第二步袁每个基因的词发生值与基值做比较遥选取词的发生值与基值的差异值超过阈值 =t+渊/n冤的词遥其中 t 是最小的阈值袁 是常数袁 是给定基因的相关摘要的数目田和 k 是主观设定的而且直接影响结果和噪声水平袁本文选择 t=15%和 k=1.5 渊这样当文献只有 5 篇时袁阈值为 45%袁而当文献数目很大时袁最小阈值为 15%冤遥第三步袁只有当至少两个基因包含同一个词时袁该词才可以用来定义基因之间的关系袁因此只有至少通过两个基因的过滤的词才保留下来遥第四步袁当噪音词太多时袁手动删除不相关的词遥

另外袁为了能挑出已经进行过大肠癌相关研究的基因袁我们在词列表中加入野大肠的冶和包括院olorectal, colon 等冤这个词袁并在生成的矩阵中袁对这个词的频率值乘以很高的权重袁这样这些基因就会聚成一类遥

1.5 等级聚类

由于词及其发生值就象微阵列的基因及其表达值一样袁因此可以基因表达谱聚类分析的方法和软件可以用来对词进行聚类分析院首先袁用经过几轮过滤后定义的词来构建一个相对于各个基因的词发生值的词 / 基因阵列遥其次袁采用 Eisen 实验室的聚类软件(Cluster)和树观看软件(Treeview)遥最后袁用 Cluster 的平均连锁等级聚类算法进行聚类分析袁用 Treeview 观看结果遥

2 结果

2.1 上调基因的聚类结果

上调基因的聚类结果见图 1遥图左边代表树结构袁上边代表用到的关键词袁右边代表对应的基因田图中黄色亮点代表该词在对应基因的相关文献中出现的频率袁最亮为频率 >20%遥图中已经进行过大肠癌相关研究的基因被分为一组袁用红色框 a 显示田在未进行过大肠癌相关研究的基因中与转移相关的基因也用红色框 b 和 c 标出遥其中袁b 和 c 框对应的词是野过表达冶野结肠直肠的冶野癌冶和野转移的冶野转移冶遥

图 1 显示 S100B袁GF4袁4.4A袁IAM1 这 4 个基因在转移的肿瘤中高表达袁但未见大肠癌相关研究遥因此袁我们更仔细地检索这 4 个基因的所有相关文献渊因前面搜索的文献不很全面袁可能出现假阳性冤袁结果发现 S100B尧GF4尧4.4A 这 3 个基因已进行过极少的相关研究袁而 TIAM1 则未发现有任何相关研究遥因此袁IAM1 值得进一步研究遥

2.2 下调基因的聚类结果

下调基因的聚类结果见图 2袁大意同图 1, 其中袁b 框对应的词是 野结肠直肠的冶和 c 框对应的词是 野相互作用冶野粘附冶和 c 框对应的词是 野转移冶遥图中显示 PAI2尧ITGAV 尧NME3 这三个基因在转移的肿瘤中低表达袁但在大肠癌中未进行过相关研究遥因此袁我们更仔细地检索这 3 个基因的所有相关文献袁结果发现 PAI2 和 ITGAV 两个基因已证实与大肠癌转移相关遥NME3 则是在结肠癌细胞株中表达噎噎袁未进行过大肠癌转移相关研究遥因此袁NME3 值得进一步研究遥

3 讨论

基于文献轮廓的微阵列挖掘技术主要目的在于用来指导复杂的表达数据库的解释遥该方法通过对大量的异质基因的功能聚类使数据变得可以理解遥从分析结果来看 TIAM1 和 NME3 是两个新的可能与大肠癌转移相关基因遥TIAM1渊T 细胞淋巴瘤侵润和转移 1冤正常主要表达于脑组织袁是一个 RAC 特异性的鸟嘌呤交换因子渊GEF冤袁属于 Rho GEFs 家族噎噎遥TIAM1 通过 ras 结合域与激活的 ras 结合袁以 PI3K 依赖的方式产生 Rac-GTP袁因此能直接调节 ras 激活 rac噎噎遥在 TIAM1 缺陷的老鼠中袁ras 诱导的皮肤癌生长缓慢遥TIAM1 缺陷与肿瘤起始的凋亡增加和肿瘤进展的增生抑制相关遥但袁同时 TIAM1 缺陷的老鼠的肿瘤的恶

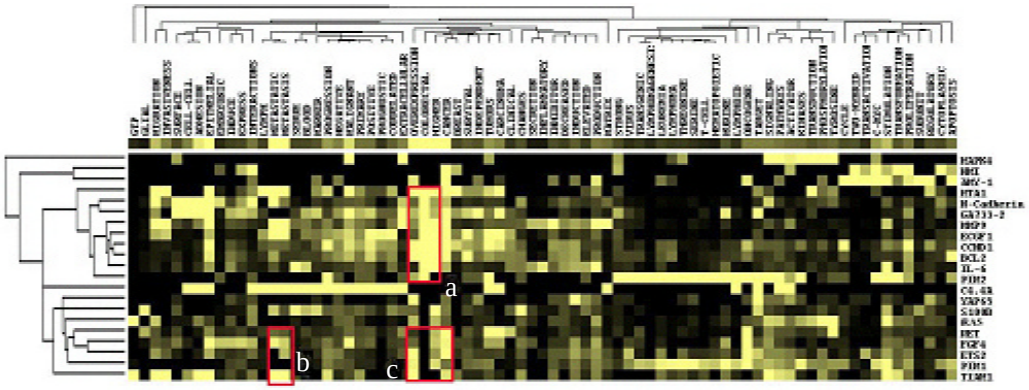

图 1　22 个上调基因的聚类结果
Fig.1　Profiling results of 22 up-regulated genes

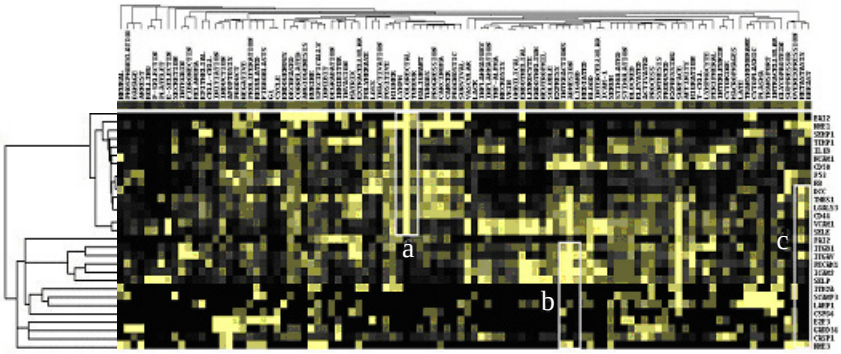

图 2　29 个下调基因的聚类结果
Fig.2　Profiling results of 29 down-regulated genes

性程度高遥因此袁TIAM1 是 ras 诱导肿瘤的关键调节物噪遥ras 基因的过表达在大肠癌的转移中起着重要的作用噪袁而且我们的基因芯片也检测到其在大肠癌转移细胞株和组织中高表达噪袁可以推测 TIAM1 基因的过表达也在大肠癌的转移中起着重要的作用遥

已知 TIAM1 与 NM23H1 相互拮抗来调控肿瘤转移噪袁NM23H1 的低表达与大肠癌转移相关噪遥我们的基因芯片也检测到 NM23H1 在大肠癌转移细胞株和组织中低表达噪遥锚蛋白渊ankyrin冤与 TIAM1 相互作用促进 RAC1 信号通信和转移的乳腺癌细胞的侵润和转移 噪0遥为了验证基因芯片的可靠性袁我们用 TIAM1 基因探针做 Northern 点杂交和原位杂交来检测大肠癌高尧低转移细胞株之间袁以及正常尧原位癌和转移癌之间的 TIAM1 表达差异袁其结果与芯片结果一致遥用 RT-PCR 检测四个大肠癌高尧低转移细胞株的 TIAM1 表达差异袁其结果也与芯片结果一致遥

NME3 渊非转移的细胞 3冤别称 DR-NM23袁是 NM23 基因家族成员袁与 NM23H1 和 NM23H2 高度同源噪遥NM23 基因家族与转移抑制和分化相关遥NME3 可能通过影响细胞外基质的粘附性和抑制软琼脂的生长来抑制成神经细胞瘤的转移噪遥结合数据挖掘尧文献分析和初步实验结果袁推测 TIAM1 和 NME3 是两个可能与大肠癌转移相关的新基因遥

参考文献院

咱暂 孙 青, 丁彦青, 高雪芹, 等. 肿瘤转移相关基因 cDNA 芯片的制备与应用咱暂第一军医大学学报, 2002, 22(12): 1070-5

Sun Q, Ding YQ, Gao XQ, et al. Development and application of cDNA microarray of tumor metastasis-associated genes咱暂 J First Mil Med Univ/Di Yi Jun Yi Da Xue Xue Bao, 2002, 22(12): 1070-5.

咱暂 Chaussabel D, Sher A. Mining microarray expression data by literature profiling咱暂Genome Biol, 2002, 3(10): 1-16

咱暂 Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images咱暂J Biomed Optics, 1997, 2(4): 364-74.

咱暂 Habets GG, van der Kammen RA, Stam JC, et al. Sequence of the human invasion-inducing TIAM1 gene, its conservation in evolution and its expression in tumor cell lines of different tissue origins 咱暂 Oncogene, 1995, 10(7): 1371-6.

咱暂 Lambert JM, Lambert QT, Reuther GW, et al. Tiam1 mediates Ras activation of Rac by a PI(3)K-independent mechanism咱暂Nat Cell Biol, 2002, 4(8): 621-5.

咱暂 Malliri A, van der Kammen RA, Clark K, et al. Mice deficient in the Rac activator Tiam1 are resistant to Ras-induced skin tumours 咱暂 Nature, 2002, 417(6891): 867-71.

咱暂 Sun XF, Ekberg H, Zhang H, et al. Overexpression of ras is an independent prognostic factor in colorectal adenocarcinoma咱暂APMIS, 1998, 106(6): 657-64.

咱暂 Otsuki Y, Tanaka M, Yoshii S, et al. Tumor metastasis suppressor nm23H1 regulates Rac1 GTPase by interaction with Tiam1咱暂Proc Natl Acad Sci USA, 2001, 98(8): 4385-90.

咱暂 Amendola R, Martinez R, Negroni A, et al. DR-nm23 expression affects neuroblastoma cell differentiation, integrin expression, and adhesion characteristics咱暂Med Pediatr Oncol, 2001, 36(1): 93-6.

咱0暂 Bourguignon LY, Zhu H, Shao L, et al. Ankyrin-Tiam1 interaction promotes Rac1 signaling and metastatic breast tumor cell invasion and migration咱暂J Cell Biol, 2000, 150(1): 177-91.

咱1暂Martinez R, Venturelli D, Perrotti D, et al. Gene structure, promoter activity, and chromosomal location of the DR-nm23 gene, a related member of the nm23 gene family咱暂Cancer Res, 1997, 57(6): 1180-7.